European Scientific Journal October 2019 edition Vol.15, No.30 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431

Principal Component and Biplot Analysis in the Agro-industrial Characteristics of *Anacardium* spp.

Maria Clideana Cabral Maia, PhD
Embrapa Agroindústria Tropical, Brazil
Adriano da Silva Almeida, PhD
State University of Piauí,
Department of Phytotechnology, Parnaíba, PI - Brazil
Lucio Borges de Araujo, PhD
School of Mathematics, Federal University of Uberlândia, Brazil
Carlos Tadeu dos Santos Dias, PhD
"Luiz de Queiroz" School of Agriculture/ University of São Paulo, Brazil
Luís Cláudio de Oliveira, Msc
Embrapa Acre, Brazil
Gilberto Ken Iti Yokomizo, PhD
Embrapa Amapá, Brazil
Renato Domiciano Silva Rosado, PhD
Department of Statistics, Federal University of Viçosa, Brazil
Cosme Damião Cruz, PhD
Department of General Biology, Federal University of Viçosa, Brazil
Lúcio Flavo Lopes Vasconcelos, PhD
Embrapa Meio Norte, Brazil
Paulo Sarmanho da Costa Lima, PhD
Embrapa Meio Norte, Brazil
Luciano Medina Macedo, PhD
Department of Plant Genetics and Improvement, Federal Technological University of Paraná, UFPR, Brazil

Doi:10.19044/esj.2019.v15n30p21 URL:http://dx.doi.org/10.19044/esj.2019.v15n30p21

Abstract

The cajuí *Anacardium* spp., which is similar to the caju *Anacardium Ocidentale* L., is a species adapted to edaphic-climatic conditions of the biome Cerratinga (Cerrado e Caatinga). Its fruit is composed of one swollen stalk (pedicel) which is formed by nutritional reserves rich in vitamin C and drupe (cashew nut). It is also rich in protein and lipids, but with smaller size. This paper focuses on investigating the applicability of the biplot graphical analysis in the process of selective breeding of cajuí population. The cajuí working

population in Embrapa Meio Norte comprises of 11 genotypes collected in areas of natural habitat in the state of Piaui. The experiment was designed in randomized complete blocks with two plants per plot and four replications. A graphical analysis (biplot) was used to study the relationships between variables and behavior of the experimental genotypes. This was implemented to principal component analysis based on singular value decomposition biplot. The total variable weight can be predicted from length of peduncle, basal and apical diameter of peduncle, and variables of easy mensuration. Genotypes M40A, M23, M14, and M17 are similar to each other and they have high amounts of brown, apical and basal diameter of the peduncle, total weight, and peduncle length. They are considered as candidates selected for consumption in natura and industrial processing. The graphical analysis (biplot) showed robustness in the presentation of relationships between variables considered and the indication of the selection candidate genotypes in the population studied.

Keywords: Native species, Biometry, Phenotypic correlations, Multivariate analysis

Introduction

Anacardium spp., known locally as *cajuizeiro*, has a small yellowish or red fruit (pseudofruit) with a fleshy, succulent pulp. The fruit peduncle is a good source of dietary fibre, both soluble and insoluble, and it presents high levels of ascorbic acid and reducing sugars. It also has high levels of vitamin A and mineral salts, such as calcium, iron, and phosphorus. Cultivating the *cajuzeiro* is an important socio-economic activity in the Northeast of Brazil, especially in the states of Ceará, Piauí, and Rio Grande do Norte where native populations of the species can be found (Carbib *et al*., 2013).

Despite being rich in nutrients and flavour and offering the native population various ways of employing the pseudofruit (peduncle) and nut in areas where it occurs naturally, the fruit of the *cajuizeiro* continues to be marketed informally and suffers extractive exploitation. It is also not included in official consumer or marketing statistics.

The *cajuizeiro* shows great potential as a source of genes for the genetic improvement of cultivated species that belong to the same genus. This is seen in the case of *Anacardium* which is used to improve the cashew (*Anacardium occidentale* L.). Selecting productive plants which show the quality characteristics of desirable peduncles, such as firmness, a high sugar and vitamin C content, and low astringency is important for defining standards and exploiting the genetic variability of the species (Rufino *et al*., 2008a, 2008b).

Multivariate analysis has routinely proven to be widely applicable to programs of plant improvement. Furthermore, biplot analysis is an efficient statistical tool for defining patterns in the relationships between variables and genotypes. It also classifies genotypes which aid in the selection of parents and superior individuals of a breeding population with the accuracy expected by the researcher.

Johnson and Wichern (1998) summarised the main techniques of multivariate analysis. They stated that among these, the technique of Principal Component Analysis (PCA) is widely used in the study of covariance structure or the correlation between variables. With PCA, a biplot can be set up to represent the variables and observations (genotypes) as a graph, thereby allowing the correlations and associations to be visualised to facilitate interpretation of the results and consequently aid in the selection of genotypes (Johnson, Wichern, 1998).

The genetic improvement of fruit trees seeks one or more homogeneous genotypes to be recommended to producers with both desirable agro-technological characteristics in their genome and repeatability of the target variables of the improvement program. The *cajuizeiro* offers the possibility of cloning genotypes at any stage of the program. This allows the selection and recombination of superior parent individuals from the base population or even the *a priori* release of elite clone(s). As such, simultaneous selection for quantitative characteristics can be facilitated using simple correlations, interrelations between the variables, and the indication of superior genotypes considered during the study. Thus, this increases the efficiency of the selection process and allows the indirect selection of more complex characteristics.

The purpose of this paper was to study the applicability of Principal Component Analysis (PCA) and biplot analysis in the selection process for improving the population of the *cajuizeiro*. This was done with a view to describe genotypes and variables, as well as the relationship between these genotypes and variables.

Materials and Methods

The working population of *cajuizeiro* at Embrapa Meio Norte comprises of 11 genotypes (M2, M12, M14, M17, M21, M23, M27, M28, M33, M40L and M40A) collected from areas where the species occurs naturally in the state of Piauí. The experiment was designed in completely randomised blocks with two plants per plot and four replications.

The variables, evaluated in a random sample of 10 fruit per genotype after reaching full physiological maturity, were coded as follows: V1 - total weight of the fruit (g), being the sum of V2 and V3; V2 - weight of the peduncle (g); V3 - weight of the nut (g); V4 - basal diameter of the peduncle

(mm); V5 - apical diameter of the peduncle (mm); V6 - length of the peduncle (mm); V7 - firmness of the peduncle (dimensionless); V8 - vitamin C (mg/100g); V9 - total soluble solids - TSS (°Brix); V10 - pH; V11 - acidity (%); V12 - total soluble solids/acidity (dimensionless).

An exploratory analysis of the data was initially carried out based on the Pearson correlation coefficient to verify the existence or nonexistence of any correlation between the variables under study. This initial analysis was followed by Principal Component Analysis (PCA). PCA is a technique of multivariate analysis that consists of explaining a complicated variance and covariance structure of a set of variables. This is done by means of few of their linear combinations in order to reduce their dimensionality and to facilitate interpretation of their interdependence (Johnson & Wichern, 1998).

Based on the PCA, it is possible to construct a biplot that represents, in two dimensions, the variables (physical and chemical) and the individuals (genotypes) (Gabriel, 1971). When constructing the biplot, an approximation (**Y**) of component 2 to the data matrix (**X**) is sought based on singular value decomposition. This approximation **Y** is decomposed as the product of two matrices **GH'**. Here, **G** is a dimension matrix (n x 2) and **H** is a dimension matrix (p x 2). The lines of matrix **G** are two-dimensional markers for the genotypes (X-line markers), and the **H'** columns represent markers for the variables (X-column markers). From these markers, it is possible to visually verify the position of one observation which is relative to another, as well as the importance of each variable for each genotype. It is also possible to verify how the genotypes and variables are grouped together. More details can be found in the study of Gower and Hand (1996).

For the results not to be influenced by the magnitude of the units of each variable, the correlation matrix was used to obtain the principal components (Barroso & Arte, 2003).

The principal components and the biplots were obtained using the BIPLOT add-on for the Excel spreadsheet software as proposed by Lipkovich and Smith (2002).

Results and Discussion

The genetic or phenotypic characterisation of genetic populations of any species is unique and nontransferable to other populations. Therefore, any comparison between the results of genetic studies, even when using a similar mathematical approach, is irrelevant and misleading. More so, they cannot be extrapolated or juxtaposed since they are the end product of the manifestation of particular genes. In addition, they are determined by varying levels of intrinsic influence from the growing environment depending on the degree of heritability of their variables. Therefore, for the effect of comparison, results obtained with other genetic populations were not reported in this study. An

exception can be justified when the analysis is based on monogenic variables (Maia *et al*., 2018). As a result, it was necessary to restrict this study to citing theoretical information and inferences.

From Table 1, it can be seen that there is a positive and highly significant relationship between the following characteristics: total fruit weight and weight of the peduncle, and total fruit weight and basal diameter Apical diameter and length of the peduncle had a negative relationship with total soluble solids (TSS). This means that fruit weight is associated with the other dimensional variables of the fruit through simple genetic control.

Table 1. Correlation matrix of physical and chemical variables of the fruits of cajuí

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	1,0000	**0,9989**	0,5476	**0,9149**	**0,6712**	**0,8487**	-0,1240	0,1890	**-0,7629**	0,4658	-0,2614	0,1870
X2		1,0000	0,5075	**0,9003**	**0,6508**	**0,8559**	-0,1328	0,1732	**-0,7629**	0,4595	-0,2420	0,1696
X3			1,0000	**0,7408**	**0,7129**	0,3213	0,0949	0,3823	-0,3996	0,3603	-0,4849	0,4105
X4				1,0000	**0,8592**	**0,6420**	0,0308	0,3174	**-0,6357**	0,5609	-0,4530	0,3742
X5					1,0000	0,3349	-0,1219	0,2918	-0,5820	0,3783	-0,3211	0,2579
X6						1,0000	-0,2805	0,3297	-0,5938	0,5264	-0,2689	0,1965
X7							1,0000	-0,0878	0,2747	0,3970	-0,6016	**0,6654**
X8								1,0000	0,2507	0,5992	-0,5789	0,4474
X9									1,0000	-0,0992	-0,0635	0,0792
X10										1,0000	**-0,9255**	**0,8844**
X11											1,0000	**-0,9751**
X12												1,0000

X1: fruit mass (FM in g), X2: peduncle mass (PM in g), X3: chestnut mass (NM in g), X4: peduncle basal diameter (PBD in cm), X5: peduncle apical diameter (PAD in cm), X6: peduncle length (PL in cm), X7: hardness, X8: content of vitamin C (mg/100g); X9: total soluble solids (TSS in %); X10: pH of endosperm (pH); X11: total titratable acidity (TTA in %); and X12: relation TSS/TTA. Values in bold: significant at 5% probability.

By measuring the length, basal diameter and apical diameter, inferences can be made concerning the total weight. This is because this characteristic is of more complex genetic control, and it can be greatly influenced by unsystematic factors.

Peduncle weight showed the same behaviour as total fruit weight in terms of correlations with other characteristics. Nut weight is significantly related to the basal diameter and apical diameter of the peduncle. On the other hand, the basal diameter of the peduncle is positively related to the apical diameter and length, but it is negatively related to the total soluble solids (TSS).

Firmness shows a significant association with the total soluble solids/acidity ratio. The pH correlates negatively with acidity, where an inversely proportional behaviour between these variables is expected, and correlates positively with the total soluble solids/acidity ratio.

The TSS/acidity ratio correlates positively and significantly with firmness and pH, but it correlates negatively with acidity.

The pH can be used for indirect inferences about acidity and the TSS/acidity ratio since it is easier to measure. In contrast, vitamin C is not related to any variable. This indicates that simultaneous selection to increase this variable may be difficult in the selection process. In addition, it can be inferred that more acidic fruits also have a lower pH.

Principal component analysis of the correlation matrix is considered when reducing the dimensionality of the interrelated variables. Table 2 shows that component 1 explains 65.34% of the total variability with 17.60% explained by component 2. Furthermore, they explain a significant percentage (82.94%) of the total standard variation (free of stochastic effects). It can also be seen from the modulus of the coefficients of the principal components that component 1 is associated with total weight, weight of the peduncle and nut, basal and apical diameter of the peduncle, and peduncle length. On the other hand, component 2 is linked to firmness.

Table 2. Component 1 (IPCA 1) and 2 (IPCA 2) for the physical variables and assessed genotypes

Variables	IPCA1	IPCA 2
Fruit mass	0,45175	-0,11915
Peduncle mass	0,44639	-0,14683
Chestnut mass	0,33317	0,42905
Peduncle basal diameter	0,45316	0,17724
Diâmetro Apical do Pedúnculo	0,37788	0,24218
Comprimento do Pedúnculo	0,36573	-0,41400
Firmeza	-0,06109	0,72020
Genotypes		
M2	-0,00574	-0,03248
M12	-0.20436	-0,09227
M14	0,23835	0,03953
M17	0,46795	-0,71142
M21	-0,32022	0,26781
M23	0,22546	0,33999
M27	-0,41364	-0,17286
M28	-0,42821	-0,10581
M33	0,06089	-0,08318
M40L	-0,03255	0,05650
M40A	0,41206	0,49419
Eigenvalue	4,57	1,23
Var (%)	65,34	17,60
Cumulative variance (%)	65,34	82,94

Component 1 shows a greater contribution of genotypes M17, M21, M27, M28, and M40A. For component 2, the following genotypes are related: M17, M21, M23, and M40A (Table 2).

In addition to the statistical data, it is important to note that one of the most well-developed and acute faculties of the human brain is the ability to perceive, analyse, and interpret complex visual information such as graphs. Patterns or visual relationships are the easiest to understand quickly, while tabulated data needs to be processed in more detail. As stated by Falk, visual information can be understood far better and much faster than linear or tabular numerical information (Yan & Kang, 2003). This is where the biplot becomes an important tool to aid the breeder by complementing and ratifying results obtained numerically.

Physical characteristics (Figure 1) falling into different sectors of the biplot indicates that their interrelationship is negative. Although when present in the same sectors, the interrelationship is positive. Based on these relationships, firmness is independent of the other characteristics and correlates negatively with length, total weight, and peduncle. Visually, total weight and peduncle had a similar contribution. This is also same for apical and basal diameter.

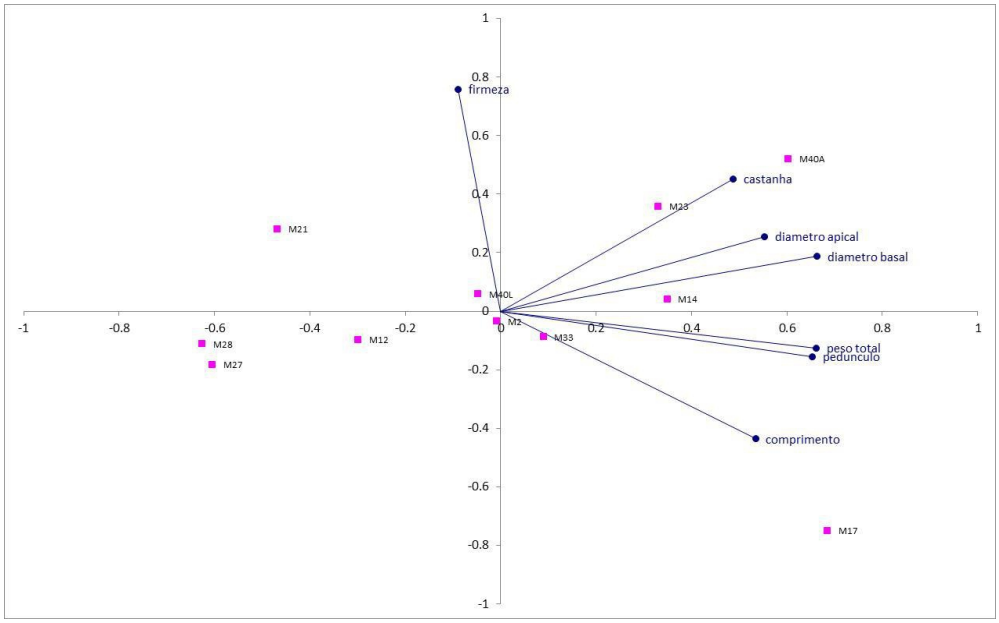

Figure 1. Biplot with two components (IPCA 1 X axis and IPCA 2 Y axis) for the physical variables and genotypes

Basal diameter presented the greatest weight, and this contributed most to the first principal component. This is followed by total weight and peduncle weight with characteristics that allow a better visualisation of any differences that may be present. Firmness stood out along the second component axis.

Genotypes M17 and M40A, with longer lines, contributed most to the observed variations. The first correlated with length and the second correlated with the nut. This is followed by M23 (with the nut) and M14.

It can be seen from Figure 1 that firmness has no relationship to any dimensional variable of the fruit. There is, however, a relationship between the other variables as shown in Table 1. Genotypes M12, M21, M27, and M28 have lower values for nut and peduncle weight, apical and basal diameter, total weight, and peduncle length. However, genotypes M40L, M2, and M33 present intermediate values for these variables. As a result of this, they are not included in the selective breeding process.

Genotypes M40A, M23, M14, and M17 make a major contribution to the genetic variability expressed by the variables, nut, apical diameter of the peduncle, basal diameter of the peduncle, total weight, weight of the peduncle, and peduncle length. This makes them good candidates for selection with a view to the industrial products market. These genotypes may be used directly in commercial cultivation through vegetative reproduction, or they may be candidates for parent plants in a progressive breeding program for the *cajuizeiro*. This is because their variability is likely to yield future genetic gains. It should be noted that genotypes M17 and M40A had the greatest contribution and there are indications that their selection would result in greater genetic gains in the population.

The principal component analysis of the correlation matrix to reduce the dimensionality of the chemical variables is shown in Table 3. It should be noted that component 1 explains 65.11% of the total variability and component 2 explains 22.08%. They both explain a percentage of the total variation of 87.19%. This is a significant amount which is able to pick up variations of genetic origin and discard the effects of noise that might make selection difficult as seen with the physical variables. This may be due to the high number of repetitions of random fruit samples for each analysis.

Table 3. Component 1 (IPCA 1) and 2 (IPCA 2) for chemical variables and assessed genotypes

Variables	IPCA1	IPCA 2
content of vitamin C (mg/100g)	0,38908	0,34467
total soluble solids (°Brix)	0,05360	0,91195
pH of endosperm	0,52822	-0.19467
total titratable acidity	-0,54334	0,06489
relation TSS/TTA	0,52105	-0,08617
Genotypes		
M2	0,41942	0,25855
M12	-0,30408	0,04989

M14	-0,36259	-0,28715
M17	-0,02563	-0,27175
M21	0,06146	0,48513
M23	0,31521	0,29284
M27	-0,29081	0,07438
M28	-0,33026	0,20086
M33	-0,20342	-0,06438
M40L	0,38865	-0,11363
M40A	0,33206	-0,62473
Eigenvalue	3,25	1,10
Var (%)	65,11	22,08
Cumulative variance (%)	65.11	87,19

In relation to the chemical variables, it can also be seen that principal component 1 is associated with pH, the total soluble solids ratio (TSS/acidity), and vitamin C. On the other hand, component 2 is linked to vitamin C and total soluble solids. In the first component, genotypes M2, M23, M40A, and M40L stand out for their contribution. In the second component, genotypes M2, M21, M23, and M28 are the most important.

Figure 2 shows that the variable, total soluble solids (TSS), does not interrelate with the other variables as seen in Table 1. Also, acidity interrelates negatively with vitamin C, TSS/acidity, and pH. Furthermore, there is a positive relationship between Vitamin C, SS/acidity, and pH. This means that a change in any one of these variables directly affects the others.

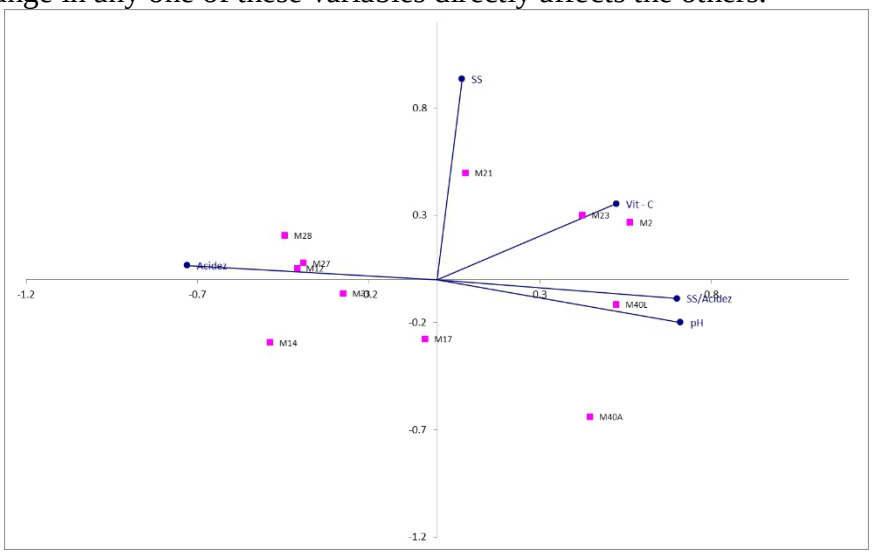

Figure 2. Biplot with two components (IPCA 1 X axis and IPCA 2 Y axis) for chemical variables and genotypes

As genotypes M23, M2 and M40L make the greatest contribution, they show high values for vitamin C, total soluble solids/acidity ratio, and pH with low acidity. Therefore, they are considered as candidates selected for *in natura* consumption. Genotype M40A has a high total soluble solids/acidity ratio and a high pH with low levels of total soluble solids and acidity. Genotype M21 has high levels of total soluble solids, and it is the major contributor to this characteristic. These four genotypes (M23, M2, M40L and M21), by contributing positively to these characteristics, show that they have the most variability available to the selection process. Also, they are the most promising for improving the population. The remaining genotypes did not stand out positively in their contribution to the characteristics under evaluation and can be ignored when selecting for improvements of the cajuizeiro.

Conclusion

Total weight can be predicted from the length, basal diameter and apical diameter of the peduncle. This is because they are easy to measure.

Vitamin C is not related to any other variable, which indicates that simultaneous selection to increase levels may be difficult.

Genotypes M40A, M23, M14, and M17 are similar with high values for the nut, apical and basal diameter of the peduncle, total weight, and peduncle weight and length. They can be considered as candidates selected for both artisanal and industrial processing.

Using graphical analysis (biplot), it was possible to select more divergent genotypes associated with physical and chemical variables.

Conflict of Interests

The authors have not declared any conflict of interests.

References:

1. Carnib, L. P. A., Aguiar, A. O., Oliveira, B. B. R., & Moreira-Araújo, R. S. R. (2013). Características físico-químicas, conteúdo de nutrientes e fenólicos totais no cajuí (Anacardium humile). Nutrile. v. 38, n. suplemento. p. 206-206.
2. Gabriel, K. R. (1971).The biplot graphic display of matrices with application to principal component analysis. Biometrika, v. 58, p. 453-467.
3. Gower, J. C. & Hand, D. J. (1996). Biplots, London: Chapman and Hall, 277p.
4. Johnson, R.A. & Wichern, D.W. (1998). Applied multivariate statistical analysis. Madison: Prentice Hall International, 816p.

5. Lipkovich, I. & Smith, E. P. (2002). Biplot and singular value decomposition macros for Excel. Journal of Statistical Software, v.5, p.1-15.
6. Maia, M. C. C., Araujo, M. F. C., Araújo, L. B. De., Dias, C. T. dos S., Oliveira, L. C., Cruz, C. D., Vasconcelos, L. F. L., Macedo, L. M., Yokomizo, G. K-I., & Lima, P. S. C. (2018). Genetic Divergence Among a Breeding Population of Hancornia Speciosa Gomes (Mangabeira) as Determined by Multivariate Statistical Methods. European scientific journal, v. 14, p. 421-433.
7. Rufino, M.S.M., Corrêa, M.P.F., Alves, R.E., & Leite, L.A.S. (2008). Utilização atual do cajuí nativo da vegetação litorânea do Piauí, Brasil. Interamerican Society for Tropical Horticulture, v.52, p.147-159.
8. Rufino, M. S. M., Alves, R. E., Aragão, F. A. S., Vasconcelos, L. F. L., Corrêa, M. P. F., & Soares, E. B. (2008). Análise multivariada de genótipos de cajuizeiro em áreas nativas da Região Meio-Norte do Brasil. Interamerican Society for Tropical Horticulture, v.52, p.140-143.
9. Yan, W. & Kang, M.S. (2003). GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists. CRC Press, BocaRaton, FL, 271 p.